Spoken Corpus of Radiotelephony Phraseology

MIRA PAVLINOVIĆ
Department of Aeronautics
University of Zagreb, Faculty of Transport and Traffic Sciences
Vukelićeva 4, Zagreb
CROATIA
mira.pavlinovic@fpz.unizg.hr www.fpz.unizg.hr
DAMIR BORAS, Phd.
Department of Information and Communication Sciences
University of Zagreb, Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb
CROATIA
dboras@ffzg.hr www.ffzg.unizg.hr
BILJANA JURIČIĆ
Department of Aeronautics
University of Zagreb, Faculty of Transport and Traffic Sciences
Vukelićeva 4, Zagreb
CROATIA
biljana.juricic@fpz.hr www.fpz.unizg.hr

Abstract: - Voice communication is one of the vital parts of air traffic control. It is "the eyes and the ears" of every pilot. It helps pilots and air traffic controllers operate the plane and maintain safe and expeditious flight. Throughout the years, investigations of many accidents and incidents have found that lack of radiotelephony knowledge and discipline by pilots and controllers has been a major causal factor. According to research, on average one miscommunication happens every hour per radio frequency. This paper gives an outline of a radiotelephony communication spoken corpus that was compiled and will be used as a basis for designing a language technology system that should spot deviations from the prescribed usage of radiotelephony communication. In order to build a corpus there are a number of factors which need to be taken into consideration. These include size, content, representativeness and usability and are discussed in this paper.

Key-Words: - radiotelephony communication, spoken radiotelephony communication corpus, aviation, air traffic control communication, approach and tower control, language technology system

1 Introduction

Aviation is a high risk environment. The task of a pilot and controller is to perform a safe and expeditious flight. During the whole flight, from starting an engine to landing and parking a plane, pilots and controllers are obliged to maintain efficient and effective communication.

During the last fifty years aviation has seen many advances. Many new systems have been implemented and many regulations introduced. The only segment that has been neglected is communication system between air traffic controller and pilot. Except for the introduction of data link, nothing has been done in this field.

The survey conducted by the NASA Aviation Safety Reporting System revealed that 80 % of incidents or accidents are attributed to incorrect or incomplete pilot - controller communication. To maintain the highest level of safety, International Civil Aviation Agency (ICAO) prescribed strict rules that govern communication between a pilot and controller. The rules for this language, radiotelephony phraseology, are located in Annex 10, Volume II, and Chapter 12 of Doc 4444 and further explained and implemented by national service providers. In Croatia this is done by Croatia Control Ltd. It sets the communication system architecture that provides fast, safe and reliable flow of information between aircraft in the controlled airspace and Air Traffic Control (ATC) centres as well as between Croatian and foreign ATC centres. As ICAO standardized phraseology is not fully harmonized on a worldwide basis every states publish differences with respect to ICAO Standards. Croatia Control Ltd., Aeronautical

ISBN: 978-960-474-304-9

Information Service, issues *Radio Communication Procedures (Voice Communication in Aeronautical Mobile Service)* in a document called AIC. The Croatian radiotelephony phraseology, technique, and procedures are based on ICAO (Standards and Recommended Practices). This paper gives an outline of the spoken radiotelephony communication corpus that was compiled and created as a part of a doctoral study research. It describes the corpus, presents its main advantages, constrains and possibilities for further research and application.

2 Radiotelephony phraseology

Radiotelephony phraseology provides means by which pilots and ground personnel communicate with each other. It is a set of prescribed rules what to say, how to say, when to say something, and how to understand uttered. Radiotelephony phraseology is an organised system for transmission of information, advice, instructions, clearances and permissions from the sender to the receiver and vice versa. It is also important to acknowledge that radiotelephony phraseology represents a set of operational procedures.

It is carried out in English, but it differs a lot from general English. It is a restricted and coded sublanguage with reduced vocabulary in which each word has a precise meaning that is often exclusive to the aviation domain. Sentences are short, determiners (the, your, etc.), auxiliary verbs (can, could, may, etc.), link verbs (is, are), subject pronouns (I, we, you, they, etc.) and many prepositions are removed. Around 50 per cent of sentences are in passive or imperative form. Here are some examples of radiotelephony language:

Cleared to land.
Reduce your speed.
Taxi straight ahead.

Standard Phraseology helps lessen the ambiguities of spoken language and facilitates a common understanding among speakers.

Pilot – controller communication is characterised by a specific communication loop:

1. The controller utters an instruction or a clearance through a headset system.
2. The instruction is transmitted through a satellite network to the pilot.
3. The pilot than receives the instruction using the headset and replies back.

Pilots always have to read back instructions received from air traffic controllers and controllers have to listen to the readbacks and confirm them.

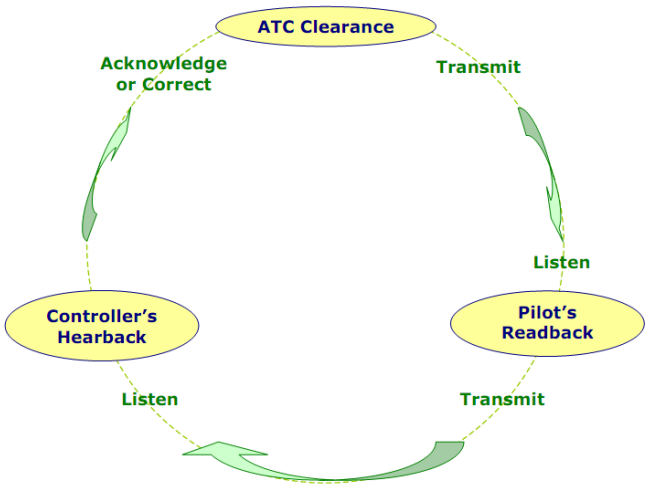


Fig. 1. Pilot – controller communication loop[5].

The items listed below have to be read back fully by the pilot. The mandatory items are:

a) Taxi/Towing Instructions
b) Level Instructions
c) Heading Instructions
d) Speed Instructions
e) Airways or Route Clearances
f) Approach Clearances
g) Runway-in-Use
h) Clearance to Enter, Land On, Take-Off On, Backtrack, Cross, or Hold Short of any Active Runway
i) Secondary Surveillance Radar Operating Instructions
j) Altimeter Settings
k) VHF Information
l) Frequency Changes
m) Type of ATS Service
n) Transition Levels [5].

If the controller does not receive the readback, the pilot is asked to do that. The pilot will request from the controller to repeat or clarify the instruction or clearance that is not fully understood.

3 Corpus collection, design and analysis

The goal of the mentioned doctorial study research is to look into communication flow within air traffic control services, and to develop and propose a language technology system that could spot deviations from the usage of standard phraseology and warn about incorrect readbacks.

2.1 Corpus collection

One of the prerequisites for setting up the language technology model was compilation of radiotelephony communication corpus. The first

ISBN: 978-960-474-304-9

idea was to compile a corpus with all instructions and clearances listed in *Radio Communication Procedures (Voice Communication in Aeronautical Mobile Service)* published by Croatia Control. After listening live communication on the frequency, it was realised that around 40% of communication differs from the prescribed communication, but even as such it is widely used and understood. So, it was decided that messages frequently used and accepted as valid by pilots and controllers in live radio communication will be included in the construction of the model. If only standardized radiotelephony were used as basis for setting the model, the model would the majority of time report on incorrect utterances and would not be functional.

Therefore, it was decided to make recordings of live radiotelephony communication, extract phrases that are frequently used and recognised as acceptable, and compile a corpus that consists of recorded phrases and prescribed radiotelephony phraseology.

The recordings used for corpus design were collected during November and December 2012 and January 2013 on the frequencies of Zagreb Approach Control (120.7 MHZ) and Zagreb Tower Control (118.3 MHZ). Icom VHF air band transceiver IC-A24, Omnidirectional Base Station Antenna CXL 3-1LW and a laptop were used for making recordings. Icom IC-A24, a device that receives and transmits radio waves for the 118 - 137 MHz civil aircraft band, reduces noise caused by atmospherical discharges was connected to outdoor base station antenna. The received signal was recorded on the laptop by a Goldwave commercial digital audio editor software and stored as mp3 files. Mp3 files do not require a lot of storage memory and are easy to handle and process.

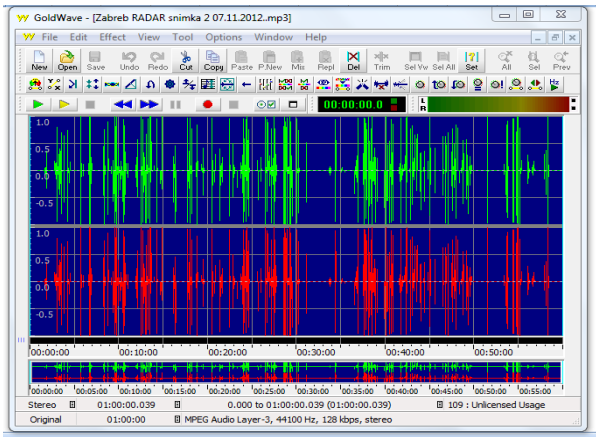


Fig. 2. Goldwave user inreface for the recording made on 7th November 2012.

Although the equipment used for corpus recording is suitable for recording in noisy surrounding and difficult conditions, some recordings were quite demanding for transcription due to bad reception, noise and occasional interruptions in the receipt of signal. It was needed for some recordings or parts of recordings to be played numerous times in order to understand communication. It took approximately five to six hours to transcribe one hour of the recorded communication.

Forty hours of communication were recorded on Zagreb Approach Control (120.7 MHZ) and Zagreb Tower Control (118.3 MHZ) frequency.

The recording were made during peak hours of traffic at Zagreb Airport Pleso, i.e. during morning hours (form 8.00 to 11.30), middle of the day (from 14.30 to 17.00), and evening hours (from 19.00 to 22.30). Taking into consideration the quality of the recordings and traffic density, out of recorded forty hours, twenty hours (ten hours of communication from Zagreb Approach Control and ten hours from communication on Tower Control) were selected to be transcribed.

2.2 Corpus design

The corpus is designed from three different groups of data. The first group consists of 556 standard radiotelephony phrases prescribed by *Radio Communication Procedures.* The second group is designed from transcripts of the recordings and contains 1967 exchanges. The number of exchanges relates to the number of massages exchanged between a pilot (P) and a controller (C). An extract from a conversation from 7th November 2012 contains five messages:

C: *Lufthansa One Papa Hotel you will be number two reduce speed to two five two five zero knots.*

P: *Reducing two fifty Lufthansa One Papa Hotel.*

C: *Lufthansa One Papa Hotel descend to flight level one zero zero.*

P: *Lufthansa One Papa Hotel descending level one hundred.*

C: *Lufthansa One Papa Hotel from present position fly on heading two two zero maintain flight level one zero zero.*

The third set of contains terminology used at airports (e.g. names of airport vehicles, names of runways and taxiways at Zagreb airport, etc.) information relevant for Croatian airspace (waypoints, routes, etc) and information on procedures that are carried out at Zagreb airport. Although the last set of data is relevant only for Croatian airspace, the users of Croatian airspace are of various nationalities and it can be stated that the

ISBN: 978-960-474-304-9

set of collected phrases is representative for radiotelephony language used in Europe.

2.3 Corpus analysis

When the recordings were collected, transcripts made, and necessary information collected, the radiotelephony corpus was designed and analysed with ***Oxford WordSmith Tools 04*** (2007). *Oxford WordSmith Tools 04* is a set of linguistic tools used for determining how words behave in a text. It consists of several tools: *WordList, Keywors and Concorde tool.* The system requirements needed for using these tools are an average computer with *Windows 2000* or later and texts saved as plain text (.txt) file.

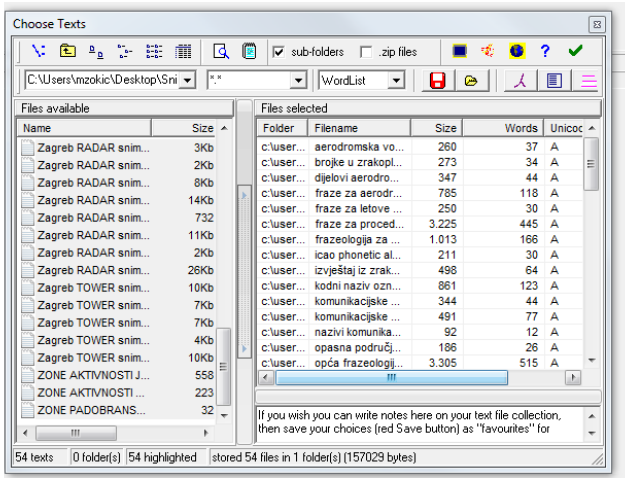
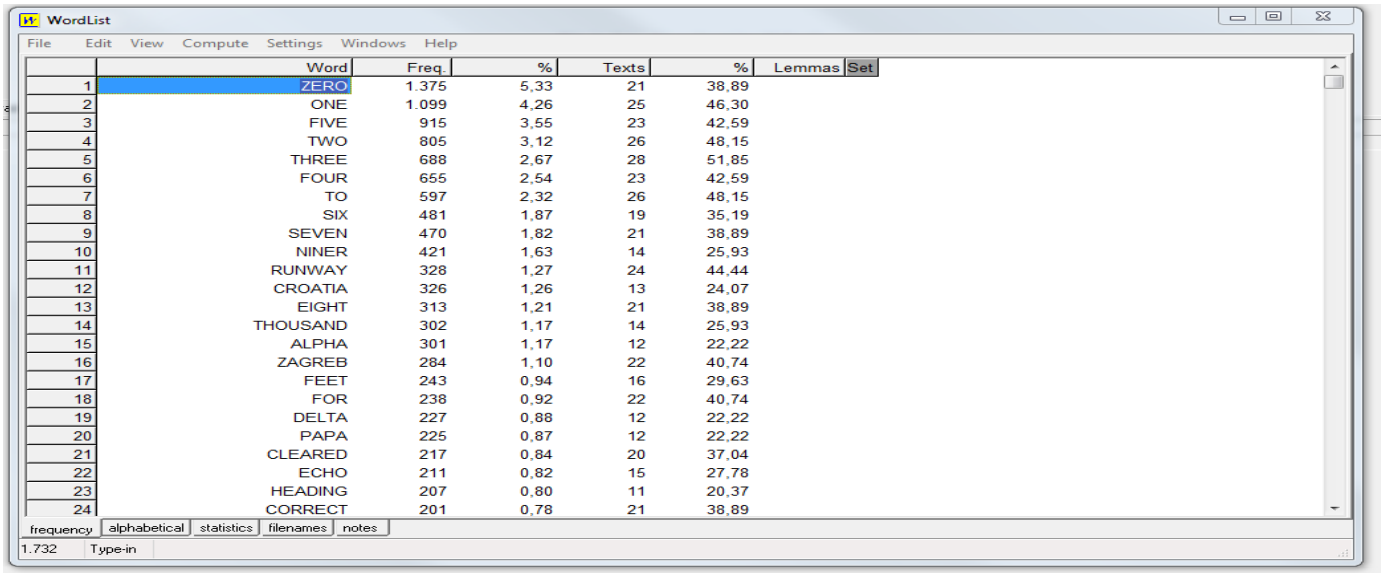


Fig. 3. A list of selected texts.

WordList tools enables us to see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order. The concordancer, *Concord*, gives us a chance to see any word or phrase in context and with key words in a text can be found with *KeyWords* tools[9].

The first step in compiling the corpus was creation of a word list. The *WordList* tool generates a list of all the words (tokens) or word forms that are included in the compiled corpus and statistical data.
It shows how often each word occurs in the text files, what is the percentage of the running words in the text, and how many text files each word was found in. The words can be listed in alphabetical order and according to the frequency (the most frequent coming first, descending to the least frequent).
54 text files that contain all previously mentioned data were selected to be included in this spoken corpus of radiotelephony phraseology.
The corpus consists of 25828 words (tokens) and 1733 distinct words (types). Type/token radio is 6,75 and mean word length is 4,91. As it can be seen in Figure 4, according to the frequency, the first ten places in the corpus are mostly reserved for numbers. The most frequent word in the corpus is *zero.* It appears 1375 times. The next one is the word *one.* The first most frequent lexical word, besides numbers and prepositions, is *runway.*

The *KeyWord* tool is a program for identifying the "key" words in one or more texts. Key words are those whose frequency is unusually high in comparison with some norm (some larger corpus; for example British National Corpus).
The program compares two pre-existing word lists, which are created using the *WordList* tool. One of these is a large word list which will act as a



	Word	Freq.	%	Texts	%	Lemmas	Set
1	ZERO	1.375	5,33	21	38,89		
2	ONE	1.099	4,26	25	46,30		
3	FIVE	915	3,55	23	42,59		
4	TWO	805	3,12	26	48,15		
5	THREE	688	2,67	28	51,85		
6	FOUR	655	2,54	23	42,59		
7	TO	597	2,32	26	48,15		
8	SIX	481	1,87	19	35,19		
9	SEVEN	470	1,82	21	38,89		
10	NINER	421	1,63	14	25,93		
11	RUNWAY	328	1,27	24	44,44		
12	CROATIA	326	1,26	13	24,07		
13	EIGHT	313	1,21	21	38,89		
14	THOUSAND	302	1,17	14	25,93		
15	ALPHA	301	1,17	12	22,22		
16	ZAGREB	284	1,10	22	40,74		
17	FEET	243	0,94	16	29,63		
18	FOR	238	0,92	22	40,74		
19	DELTA	227	0,88	12	22,22		
20	PAPA	225	0,87	12	22,22		
21	CLEARED	217	0,84	20	37,04		
22	ECHO	211	0,82	15	27,78		
23	HEADING	207	0,80	11	20,37		
24	CORRECT	201	0,78	21	38,89		

Fig. 4. A frequency listing for spoken corpus of radiotelephony phraseology.

ISBN: 978-960-474-304-9

reference file. The other is the word list based on one text which is studied. The list of key words has not been created as it is not relevant for this research.

a) It is representative. The criterion of representativeness is fulfilled by selection of the text. The corpus contains texts of the same register and content, that is text of radiotelephony

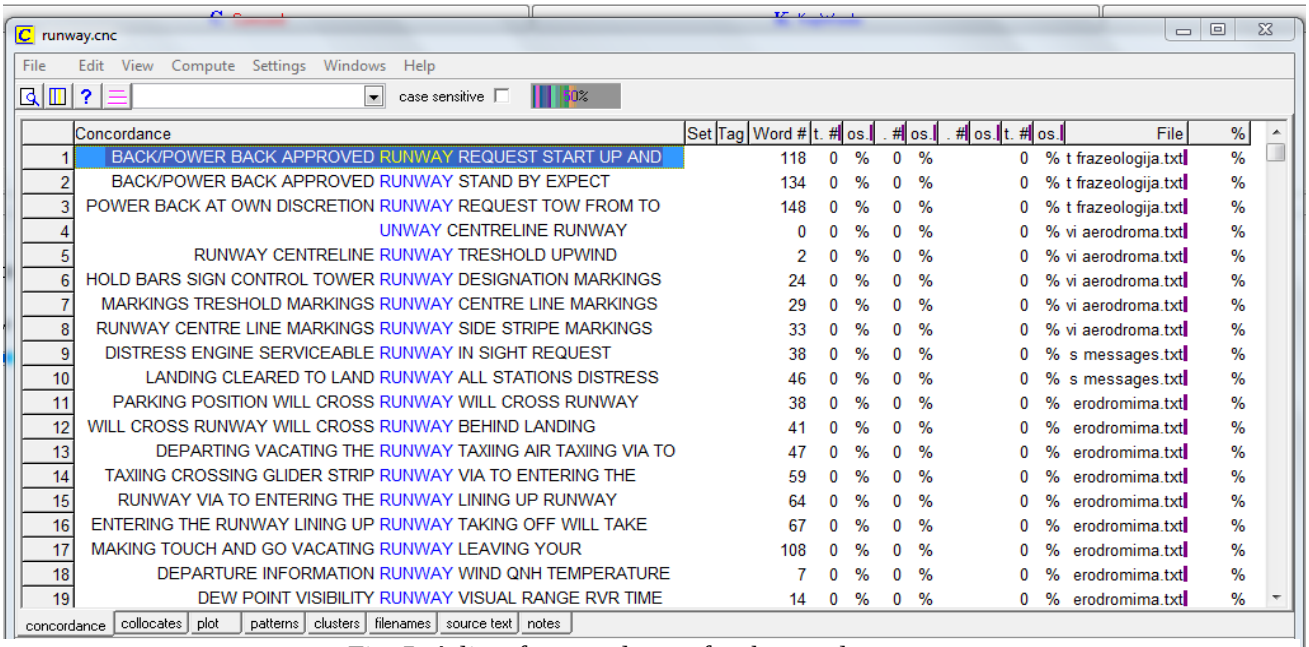


Fig. 5. A list of concordances for the word *runway*.

The *Concord* tool enables us to see lots of examples of a word or phrase in their contexts. This tool is the most important part of *WordSmith* tools for our research and has most frequently been used in designing the mentioned language technology system. The *Concord* tool has been used to make a list of phrases that differ from the standard radiotelephony phrase, but have similar meaning and are frequently used and overall accepted. The starting point have been phrases contained in *Radio Communication Procedures* (a search word or word phrase is specified). Then, the *Concord* tool looks for it in all chosen text files. And finally, the word or phrase is presented on a concordance display giving access to information about collocates and stored for further usage.

The concordances can be listed alphabetically or in the order they appear in text files. The listings can be saved for later use, edited, printed, copied to your word-processor, or saved as text files. Figure 5 shows concordances for the word *runway* and its immediate contexts. For the word *runway*, concordances are listed according to their appearance in the text files.

2.4 Features of radiotelephony communication spoken corpus

Here are some features of the designed spoken corpus of radiotelephony phraseology:

communication. The findings from the corpus are generalisable and applicable to European radiotelephony language.

b) In terms of the content, it contains standard radiotelephony phrases prescribed by *Radio Communication Procedures*, transcripts of radiotelephony communication recordings and terminology used at airports, information relevant for Croatian airspace and information on procedures that are carried out at Zagreb airport.

c) For the moment it can be described as a static corpus. We are aware that this feature may have an influence on the corpus representativeness so the plan is to extend the spoken corpus of radiotelephony communication for future research.

4 Conclusion

Although the compiled spoken corpus of radiotelephony communication has been designed for Croatian airspace, due to variety of nationalities using Croatian airspace, the designed corpus is found to be representative for radiotelephony language used in Europe and applicable to any research in European radiotelephony communications.

Even though many experts in corpus linguistics agree the larger the corpus the better, for the purpose of this research, a relatively small corpus of spoken radiotelephony language has been designed

ISBN: 978-960-474-304-9

with only 1733 words. There are two reasons for that:
1. As already mentioned, the language of radiotelephony communication is a restricted, coded and standardized sublanguage with reduced vocabulary.
2. The process of collecting materials for spoken corpus design (recording and transcription of communication) is time consuming.

Nevertheless, it has to be emphasised that smaller specialized corpora containing texts of a particular genre can be extremely useful. It is possible to get much useful data from a small corpus, particularly when investigating high frequency items, as is the case with this spoken corpus. In such corpora is easier to identify specialized terms and detect collocations, and it provides a wealth of information about structure, style and concepts in the specialized target language. All that makes concordancing more representative and utilizable.

References:

[1] AIC – Voice Communication Procedures, Croatia Control Ltd. Zagreb, 2008.

[2] Carroll, J.M., Making Use: scenario-based design of human-computer interaction. Cambridge, MIT Press, 2000.

[3] Communication Procedures including those with PANS, ICAO, Annex 10, Volume 2, 2001.

[4] Context-Sensitive Speech Recognition In The Air Traffic Control Simulation, Eurocontrol. EEC Note No. 02/2001. URL: http://137.193.200.177/ediss/schaefer-dirk/inhalt.pdf. (15.04.2012.)

[5] Effective Pilot / Controller Communications, Flight Operations Briefing Notes, Human Performance, September 2004. URL: http://www.airbus.com/fileadmin/media_gallery/files/safety_library_items/AirbusSafetyLib_-FLT_OPS-HUM_PER-SEQ04.pdf. (14.09.2011.)

[6] Juričić, B., Varešak, I., Božić, D., The Role of the Simulation Devices in Air Traffic Controller Training, Electrotechnical Association of Slovenia, ISEP 2011 Proceedings, 2011.

[7] McMillan, D., Miscommunications in Air Traffic Control, Queensland University of Technology, October 1998. URL: http://ebookbrowse.com/miscommunications-in-air-traffic-control-pdf-d350817961 (27.04.2012.)

[8] Pilot-Controller Communications, Eurocontrol. June 2004. URL: http://www.skybrary.aero/bookshelf/content/bookDetails.php?bookId=139. (13.10.2011.)

[9] Prinzo, V. O., An Analysis of Voice Communication in a Simulated Approach Control Environment, Civil Aeromedical Institute Federal Aviation Administration, May 1998. URL: http://www.hf.faa.gov/docs/508/docs/cami/9817.pdf. (03.02.2013.)

[10] Scott, M., WordSmith Tools Manual, Version 6.0. Liverpool, Lexical Analysis Software Ltd., 2013.